CLAIMS

What is claimed is:

1. A spam filtering system comprising:

one or more spam filters; and

a randomization component that obfuscates functionality of a spam filter to mitigate reverse engineering the one or more spam filters.

2. The system of claim 1, the randomization component randomizing scores of the filter so as to make it difficult for a spammer to determine whether a message that is close to a threshold and changes from being one of blocked or delivered, has changed due to one of the following: a modification to the message and the randomization component .

3. The system of claim 1, the randomization component comprising a random number generator that generates at least one of a random number and a pseudo-random number.

4. The system of claim 3, the randomization component comprising one or more input components whereby the one or more input components provide input to the random number generator to facilitate determining what random number to generate for a particular message.

5. The system of claim 1, the randomization component generating a random number based at least in part upon input received from one or more input components.

6. The system of claim 5, the input from the one or more input components is based at least in part on time.

7. The system of claim 6, wherein the random number generated depends on at least one of: a time of day and an increment of time; such that the number generated changes according to any one of: the time of day and a current increment of time.

8. The system of claim 5, the input from the one or more input components is based at least in part on at least one of: a user, a recipient, and a domain receiving the message.

9. The system of claim 8, wherein the random number generated depends on at least one of: a user, a recipient, and a domain receiving the message; such that the number generated changes according to any one of: an identity of the user, an identity of the recipient of the message, and the domain receiving the message.

10. The system of claim 9, wherein the identity of any one of the user and the recipient comprises at least one of a display name and at least a portion of an email address.

11. The system of claim 5, the input from the one or more input components is based at least in part on content of the message.

12. The system of claim 11, wherein the random number generated changes depending on at least a portion of the message content.

13. The system of claim 11, wherein a hash of the message content is computed and the hash value is used as the random number, whereby even a small change to the message content results in a substantially large change to the random number generated.

14. The system of claim 11, wherein a hash of at least a portion of features extracted from a message is computed to facilitate randomizing a message score and thus, the functionality of the spam filter.

15. The system of claim 14, wherein the features used to compute the hash have respective individual weights greater than some threshold.

16. The system of claim 11, wherein a hash of a sender's IP address is computed to facilitate randomizing message scores to thereby obscure the functionality of the spam filter.

17. The system of claim 1 having a substantial effect on messages that border between spam and non-spam, whereby messages that are border-line spam are classified as spam at least part of the time by randomizing scores of the messages.

18. The system of claim 1, the randomization component mitigating spammers from finding at least one message that gets through the spam filter substantially every time it is sent.

19. The system of claim 1, the spam filtering system making use of a sigmoid function having the formula of $finalscore = \frac{1}{1 + e^{-summedscore}}$, wherein at least one of a *summedscore* value and a *finalscore* value is randomized to effectively modify spammer behavior and to mitigate reverse engineering of the filtering system.

20. A multi-spam filter filtering system that mitigates reverse engineering of spam filters and mitigates finding one message that gets through a spam filter substantially all the time comprising:

a plurality of spam filters comprising at least a first spam filter and a second spam filter for processing and classifying messages;

a plurality of users comprising at least a first user and a second user; and

a filter selection component that selects one or more filters to be deployed for use by at least one of the plurality of users.

21. The system of claim 20, further comprising a time input component that communicates with the filter selection component such that one or more of the plurality of filters are selected and deployed for a respective user based at least in part upon any one of a time of day and a time increment.

22. The system of claim 21, wherein the time increment is any number of seconds, minutes, hours, days, weeks, months, and years.

23. The system of claim 20, the filter selection component selects the one or more filters randomly.

24. The system of claim 20, the filter selection component selects the one or more filters non-randomly.

25. The system of claim 20, the filter selection component selects the one or more filters to be deployed to the respective users based at least in part on at least one of the respective users, a domain of the sender, a domain that is operating the filtering system, and a domain receiving the messages.

26. The system of claim 20, the users being recipients of the messages.

27. The system of claim 20, wherein at least a portion of the plurality of spam filters is trained using one or more sets of training data *via* a machine learning system.

28. The system of claim 27, the training data corresponding to features extracted from messages.

29. The system of claim 28, wherein at least a portion of the features extracted from the messages is forced to have particular values.

30. The system of claim 28, wherein at least a portion of the features extracted from the messages is excluded from the training data.

31. The system of claim 28, wherein at least a portion of the features extracted from the messages is clustered by feature type such that each cluster of data is used to train individual filters

32. The system of claim 31, wherein at least a portion of the plurality of users is clustered by user type, the user type being related to the feature type clusters such that a spam filter corresponding to the user type is employed for that user.

33. The system of claim 20, wherein the first filter is trained using at least a first subset of training data and the second filter is trained using at least a second subset of training data, at least a portion of the second subset of training data being non-overlapping with at least a portion of the first subset of training data.

34. The system of claim 33, wherein the first filter and the second filter are deployed for use together so that a plurality of different criteria and/or features of the messages are looked at before classifying the messages as spam or non-spam.

35. A method that facilitates obfuscating a spam filter comprising:

running a message through a spam filter;

computing at least one score associated with the message;

randomizing the score of the message before classifying the message as spam or non-spam; and

classifying the message as spam or non-spam.

36. The method of claim 35, wherein the at least one score associated with the message comprises a *finalscore* and a *summedscore*.

37. The method of claim 36, wherein the *summedscore* is a sum of all scores associated with individual features extracted from a message.

38. The method of claim 36, wherein the *finalscore* is a sigmoid function of the *summedscore* and corresponds to a value between 0 and 1 that indicates a probability that a message is spam or not.

39. The method of claim 35, wherein randomizing the score of the message comprises adding at least one of a random number and a pseudo-random number to the score of the message.

40. The method of claim 39, the number added to the score of the message depending at least in part upon at least one of the following:

a time of day; and

a time increment.

41. The method of claim 39, the number added to the score of the message depending at least in part upon at least one of the following:

a user;

a recipient of the message;

a domain receiving the message;

a domain of the sender; and

a machine name running the filter.

42. The method of claim 39, the number added to the score of the message depending at least in part upon at least one of the following:

a hash of contents of the message; and

a hash of at least a portion of features extracted from the message.

43. The method of claim 42, wherein the features used to compute the hash have respective weights greater than 0.

44. The method of claim 42, wherein the features used to compute the hash can randomly or non-randomly change depending on at least one of a time of day and a time increment.

45. The method of claim 39, the number added to the score of the message depending at least in part upon a hash of a sender's IP address.

46. The method of claim 39, the number added to the score of the message depending on input from one or more input components.

47. A method to minimize spam comprising deploying a plurality of spam filters across a plurality of users so as to mitigate reverse engineering of the spam filters and to mitigate spammers from finding particular messages that consistently get through the filters.

48. The method of claim 47, deploying at least a portion of the plurality of spam filters depends on at least one of a time of day and a time increment.

49. The method of claim 47, deploying at least a portion of the plurality of spam filters depends on at least one or more users making use of the spam filters.

50. The method of claim 47, deploying at least a portion of the plurality of spam filters depends on at least one of a hash of message contents and a size of the message.

51. The method of claim 47, further comprising selecting at least a portion of the plurality of spam filters for deployment randomly.

52. The method of claim 47, further comprising selecting at least a portion of the plurality of spam filters for deployment non-randomly.

53. The method of claim 47, the plurality of spam filters being trained with sets of training data via machine learning processes.

54. The method of claim 53, training the spam filters comprising:

creating sets of training data;

using at least a first subset of training data to train at least a first spam filter; and

using at least a second subset of training data to train at least a second spam filter, whereby the second subset is not equivalent to the first subset of training data.

55. The method of claim 53, training the spam filters comprising:

clustering training data by type to correspond to clusters of user types:

training at least a first filter with a first cluster of data; and

training at least a second filter with a second cluster of data.

56. The method of claim 55, wherein the first filter is deployed to a user belonging to a related type of cluster.

57. A computer readable medium comprising the method of claim 35.

58. A computer readable medium comprising the method of claim 47.

59. A computer-readable medium having stored thereon the following computer executable components:

a randomization component that obfuscates functionality of a spam filter so as to hinder reverse engineering the one or more spam filters.

60. The computer-readable medium of claim 59, the randomization component randomizing scores of the filter.

61. The computer-readable medium of claim 59, the randomization component comprising a random number generator that generates at least one of a random number and a pseudo-random number.

62. A system that facilitates obfuscating a spam filter comprising:

a means for running a message through a spam filter;

a means for computing at least one score associated with the message;

a means for randomizing the score of the message before classifying the message as spam or non-spam; and

a means for classifying the message as spam or non-spam.

63. A system that minimizes spam comprising a means for deploying a plurality of spam filters across a plurality of users so as to mitigate reverse engineering of the spam filters and to mitigate spammers from finding particular messages that consistently get through the filters.